元语言研究的三种理解及释义型元语言研究评述

苏新春
(厦门大学 中文系,福建 厦门　361005)

摘要:"元语言"最早是由哲学界提出的一个命题。在后来的发展中,形成了语义哲学界与对象语言相对立的表达语言、语言学词典学中与被释词相对的释义语言及自然语言处理界中表示认知义原的三种不同的主要研究类别,并正在成为愈多的人文科学所接受的一个底层理论的概念术语。《朗曼当代英语词典》的 2000 条释词体现了英语的释义元语言工作,它形成有限释词集的是作者丰富的教学经验与词典编纂经验。国内信息处理界也进行了释义元语言的工作,依靠的是数学中的图论方法,形成了 3800 多条的"定义原语"。然其存在不足,主要在于名词特别是专用名词所占的比重过多。

关键词:元语言;释义元语言;词典学;语义学
中图分类号:H 06　H 031　　**文献标识码**:A　　**文章编号**:1000-579(2003)06-0093-10

Three Interpretations of Metalanguage and Different Researches on Defining Metalanguage

SU Xin-chun
(Xiamen University, Xiamen, Fujian 361005)

Abstract: The notion of metalanguage originates from philosophy. In the later research there come three interpretations: "expression language" as opposed to "object language" from the perspective of semantic philosophy, "defining language" as opposed to "defined language" from linguistic lexicography and "cognitive semantic primitives" from natural language processing. And now it is growing to be a fundamental concept in some fields of humanity science. There are 2000 entries based on the editors' experience on teaching and lexicography in LDCE, Wwhich stands for the work on defining metalanguage in English. The 3800 entries of defining metalanguage made by the Chinese scholars in information processing are following the graph approach in mathmatics, while there are also shortcomings such as the large proportion of nouns especially the proper nouns.

Key words: metalanguage; defining metalanguage; lexicography; semantics

一、"元语言"的提出与三个领域的不同理解

"元语言"最早是由哲学界提出的一个命题。20 世纪波兰逻辑学家塔斯基(Alfred Tarski)认为人们当判断一句话是真还是假时,往往会把这句话的客观真实性与这句话存在的真实性混淆在一起。因此,在区别语言与语言所指称的事物的关系时,就有必要把真实语言与形式语言区分开来。真实语言是与客观对象相联系的语言,在与元语言相对时称之为对象语言。而用来称说对象语言的则是元语言。自塔斯基提出元语言后,引起了人们广泛的思考,逐步扩大到了许多有关的学科,并形成了三种不同类型的元语言理论。

1. 语义哲学界的元语言说

语义哲学界是元语言理论的发源地。塔斯基关于元语言理论的理解产生了广泛影响。一本有影响的哲学著作对

收稿日期:2003 - 09 - 12
作者简介:苏新春(1953 -),男,江西南昌人,厦门大学中文系教授。

塔斯基的元语言理论作了详尽的介绍。

(塔斯基认为)对每一种有穷阶的形式化的语言来说,一个形式正确和实质适当的关于真句子的定义能够在元语言中如下构造出来:只使用一般的逻辑表达式,语言的本身的表达式,以及属于语言构词法的术语,即语言学表达式的名字和这些表达式之间存在的结构关系的名字。[1](P148)

他在论述到"X 是真的当且仅当 P"这一关于"真"的定义时,就认为这一推演形式,及这一推演形式中的 X、P 都是属于元语言的范畴。"P"指的是一个句子,"X"指的这个句子的名字。在这里的元语言中,"真"定义成立的关键依靠的是等值表达式,而不是"X"或"P"所指称的客观事实。

一部有影响的哲学辞典《西方哲学英汉对照辞典》介绍了哲学界关于元语言的一个经典解释:

> 第一语言的表达式的名称,以及这些表达式之间关系的名称,都属于第二语言,后者叫做元语言。①

哲学界的元语言是相对于真实语言而言,它没有了真实语言所包含的现实世界的那些具体、庞杂、混合的内容,而只是真实语言的抽象语言表达形式。元语言有着抽象的、形式化的、纯粹的、超越客观所指对象的形式语言特点。后来的哲学研究者对元语言的提出给予了很高的评价,认为它克服了长期以来将思想、存在、语言三者混为一谈的不足:

> 我们现在回头来看,语言之所以能成为西方哲学的最后边界,是因为西方哲学从一开始就将对于'存在'(古希腊语 on,英语 to be)的思考视为自己的核心课题,而 on 或 to be 具有双重意义:它是哲学意义上的'存在',又是语言学意义上的系词'是'。前者是从'对象性语言'层面对事实的陈述,后者是从'元语言'层面对思想的表述。在前一种情况下,它是一种对象性的陈述或描述;在后一种情况下,它是一种元语言性质的判断或断定。于是,'事实 - 思想 - 语言'打成一片了,或曰混为一谈了。西方哲学这种以'言'代'有'、以'思'代'在'的理性主义传统,确实异常强大,以至于现代人文主义最杰出的哲学大师海德格尔,最后也未能彻底逃出'语言的牢笼',以至承认'语言是存在的家园'。不仅如此,当今西方哲学似乎还有某种越陷越深的迹象。[2]

> 为了能谈论陈述,就必须使用陈述的名称或陈述的描述,也许还有像'陈述'那样的词,即理论必须用元语言、用人们用以谈论语言的语言。而为了能谈论事实和有意义的事实,就必须使用事实的名称或事实的描述,也许还有像'事实'那样的词。一旦我们有了元语言,类似这种我们能用以谈论陈述和事实的语言,就容易就陈述和事实之间的符合作出断言。[3]

哲学界的元语言理论在语言学界有着很大影响。一些有影响的语言学著作或辞典所引述的看法大都沿用了这一看法。《语言与语言学词典》对 metalanguage 的解释是:

> 元语言,纯理语言。指用来分析和描写另一种语言(被观察的语言或目的语[Object language])的语言或一套符号。[4]

有的则兼采众说,而把哲学界的看法放在首位。如《语言学百科词典》:

> 元语言,又称"纯理语言"、"符号语言"。与"对象语言"相对。指描写和分析某种语言所使用的一种语言或符号集合。用汉语来说明英语,英语是对象语言,汉语是元语言;用英语来说明英语,英语既是对象语言,又是元语言。在辞书编纂和语言教学中用于释义的语句称元语言;在语言研究中为描写和分析语言成分特征使用的一套符号和术语,如[±Noun]([±名词])、[±Abstract]([±抽象])、[±Animate]([±有生命])等,也属元语言。[5]

2. 语言学词典学中的元语言说

语言学特别是词典学所谈到的元语言虽然也受到语义哲学中元语言说的影响,但它们所指已经有了很不相同的含义,指的是用来解释词典所收词语的定义语言——本文称之为释义元语言。②

Wierzbicka 曾有过一段很精彩的论述:

> 1)任何语言的词典中都存在不可定义的词,它们的数量较少,自成系统,它们的作用是用来定义其它的词语。2)不可定义的词是可列举的,语言中的其它词可以用它们来定义。3)不可定义的词在不同的语言中虽然各有所不同,但却是相互对应的,在语义上是等价的。因此,不可定义的词在各种语言中可视为"普遍词汇"。[6](P3)

我们可以看到,定义语言的某些特点与语义哲学有着共通之处,例如它们是有限、可穷尽列举的,在不同的语言中是共通的、等价的等。但二者之间又有着很不相同的东西,关键在于,语义哲学中的元语言是脱离事实语言的,属于形式语言的一部分,它是高于事实语言,被抽象出来的语言表述格式,而释义元语言则是事实语言的一部分,是其中通用、高频、中性的那一部分。

用作解释别的语言成分的释义元语言说法,在我国语言学界、词典学界有相当大的影响。这两大学科的学者一般都是从这个角度来阐释元语言的理论。

辞书中解释词条的语言,是元语言之一。这种元语

---

① 塔斯基(Alfred Tarski,1901 - 1983),波兰逻辑学家。代表著作是 The Concept of Truth in Formalized Languages,in Logic,Semantics,Mathematics,Oxford,The Clarendon Press(1956)。

② 本文在指称整体时用"释义元语言",而在指具体词语时用"释义元词",用来释义的一般性语言称为"释义词"或"释义词语"。

言的整体观包括元语言的整体简化,即只使用民族共同语的有限的常用词。[7](P346)

所谓"元语言",是英语 metalanguage 的汉译,指的是用来分析和描述语言的语言。有人觉得"元语言"的译法不知所云,宁愿接受"纯理语言"或"前设语言"的译法。总之,这是一种"工具语言"或"人为语言",而不是日常应用的自然语言。一般词典释义就经常要运用"元语言"以及与此相关的多种符号和格式。有的词典家认为,借用"元语言"释义,令读者增加一层负担,未必有利于读者掌握词义,倒不如直接运用自然语言交待词义,更便于与读者"交流"。于是就有了前文所介绍的"不要释义"的尝试。[8](P122)

3. 自然语言处理界的元语言说

在自然语言处理界,如何让计算机能自如地处理繁复无比的语言一直是一个"引无数英雄竞折腰"的课题。而试图把语言形式化、规则化,并最终能做到自如地生成语言,则是人们探索的一条基本思路。在语法、语音的形式化、规则化完成以后,语义又放在了人们的面前,这是最难处理的一部分。要使语义做到形式化、规则化自然会产生出原始语义的想法,这就是"语义原语"的来由。Yorick Wilks 对语义原语下过这样的定义:

A "PRIMITIVE" (or rather a set of primitives plus a syntax etc.) is a reduction device which yields a semantic representation for a natural language via a translation algorithm and which is not plausibly explicated in terms of or reducible to other entities of the same type. 原语(或者说一个原语集加上一个句法)是一个语义消减装置,自然语言可能通过一个翻译算法转化成用原语进行的语义表示,而原语本身不能再消减成或解释成其它同类实体。[6](P3)

这里关于语义原语的说明显然是吸收了语义哲学中元语言的思想,如它有最小性,不能被再分解;有生成性,能够由原语再加上某些规则来做新的表示;有形式语言的特点,能够由翻译算法、代码等来指代自然语言。不同之处就是它是由计算机来认知与操作完成的。

以上三种元语言理论:语义哲学中具有形式语言特点的元语言说、词典学中的释义元语言说、自然语言处理中的语义原语说,对元语言的"元"显然有着很不相同的理解。在它们眼里,元语言各有着不同的性质和特点,不同的功能与作用。由于词典学是一门很注重实践的学科,收词释义是词典的基本内容,以解释词语为己任的释义元语言研究也就成为元语言研究中富于实践意义、具有良好的可操作性、具体性研究做得最充分的领域之一。释义元语言又是词汇学中具有基础意义的一块重要内容,在研究中要运用到词汇的性质、词语的分类与分层、常用词、词义的义域、词义的系统性等众多理论。

二、"元语言"正在成为对众多学科产生影响的普遍理论

人们早就认识到在一种语言的千千万万个词语中,它们的地位并不是完全相同的,而是存在着使用的频偶、年代的久暂、意义的广狭、影响的大小、再生的强弱、涵蕴的丰寡、识别的难易等诸多差异。这些差异会直接影响到具体词语在词汇系统中的地位。这也就是人们总是乐于对词语总汇进行分类逐层、条分缕析研究的根本原因。元语言理论的提出,就是在这样一种求知背景下做出的探索努力之一。

在当代,"元语言"已是一个相当热闹的话题。人们已经愈来愈多地把元语言看作是一种语言词汇系统中位于最核心的位置,最富于解释力,能成为其它语言成分的诠释工具的那部分语言。笔者 2003 年 6 月 29 日在"新浪网站"键入"元语言",索得资料多达 553 条,而用搜索引擎 gpogle,竟索得 1005 条。发现人们在谈论众多学科问题时已经把元语言作为一个底层理论的术语根据自己的理解来随时加以引用。

如教育心理学之对元语言:

从目前已有的研究看,汉语儿童学习字词和阅读的同时,也发展了各种元语言学意识,其中一些元语言学意识的发展与儿童阅读能力发展有很密切的关系。[9]

培养学生的"元语言能力"……谈到"教养"之基础的语文素养,使我联想到大西道雄教授的一番论述。他首先把"有教养"的人界定为这样一种人:"立足于语言感悟,能够深入地思考、准确地判断和行动自立的人",也可以说是"能够借助语言同他人交流,同时,基于相对的自我认知拥有同他人共同生存的能力的语言主体。"然后强调了构成语文素养之核心的要素,就是"思维语言"。或者模仿或者根据指令作出行动的所谓"机器人"是同这里所说的"有教养"的人风马牛不相及的存在。而使人超越了这一点的,便是"元语言思维"。[10]

"元语言意识"指儿童对语言、文字一般结构特征的认识和操作。[11]

文学创作之于元语言:

元语言在文学中已非纯粹语言学上的解释,从雅可布森到巴特,文学批评成为对文学而言的一种元语言,于是任何学科都有属于自己的元语言,到了拉康,元语言泛化、相对的意义使之取消了自身的存在,融进了后现代。我依然承认元语言的存在,如同承认文学批评的意义。元语言可以是意义生成的,也可以是对日常意义的毁灭,或者两者兼有。[12]

编程语言之于元语言:

XML 是一种结构化描述语言。它随着因特网技术和电子商务的发展成为 HTML 的后继者。它的优势在于,它不仅是一种标识语言,更是一种可以定义描述对象结构的元语言。XML 文档自含结构,使得系统间交换

的信息可以互相"理解"。[13]

根据定义,XML 文件是合乎规范的 SGML 文件,是 SGML 的一种简化形式,也是一种能够定义其他标记语言的元语言。[14]

摄影家之于元语言:

"语言学转向"对于摄影意味着,照片无论如何"逼真"地记录了现实,也不能因此便把它与现实等同起来。因为在照片与现实之间,还隔着一层并不透明的东西——符号(或曰语言,一种"光影语言"以及其"元语言")。[15]

语篇学者之于元语言:

通过对不同语篇体裁的语篇进行比较与对比,我们不仅可以从微观上把握一类语篇的内在结构及其组织机制,并能从宏观上了解某一语篇体裁发生的社会文化背景或语境。换言之,语篇体裁创造了对一类语篇进行整体描述的元语言(metalanguage)。[16]

医学者之于元语言:

医疗卫生方面的越来越多的信息需求使疾病分类问题越来越突出,似乎上个世纪的诊断词汇快速膨胀,没有相应的精确的元语言与之匹配以描述诊断术语之间的关系。尽管一些元术语如疾病、紊乱、综合征等已经被广泛使用,但其确切含义方面仍有很多模糊,描述疾病分类学关系的元语言也仍缺乏或未被应用。①

这不由得不让人深深感到,"元语言"正在走出哲学界、语言学,正在成为一个具有泛学科意义的普通术语。尽管它在不同学科有着不同的含义,但其共同特点似乎都具有了以下的意味:具有超现实的意义,不含有使用通常语言时一般会含有的指称别的事物的杂质;是对其它语言表达形式的解释者或构成者;在该学科领域中具有底层理论建构的意义等。显然,所有的这些学科对元语言的使用都是后续性的,而它却极大地受到语言学关于元语言理解的影响。

三、《朗曼当代英语词典》的释义元语言

在语言学和辞典学的范围,释义元语言成为人们关心的焦点。对释义元语言的研究,西方语言学进行得相当充分,其中又以英语为最。

《朗曼当代英语词典》用近 2000 个常用词解释 56000 个词条,威斯特和因迪科特的教学词典(第 4 版)用 1490 个词解释 24000 个词条,法国古根海姆两卷本词典元语言包括 1374 个'成分词汇'和 55 个下定义词。下定义词大约指的就是属词。[7](P346)

要说到英语词典学界的释义元语言研究,就不能不说到迈克尔·威斯特和他的《新方法英语词典》。② 威斯特毕生从事英语教育工作,早年他运用自己的心理学知识及行为主义理论模式,通过简化词汇、改变词汇分布结构、运用常用词并逐渐加入新词的方法,来设计新的阅读方法,编写新的阅读教材,取得了明显的教学效果。威斯特对阅读方法的研究导致了他对词汇控制理论的研究,当时,词汇控制研究是最热门也是最有争议的外语教育研究领域。参加讨论的四位著名学者桑代克(E.L. Thorndike.)、帕尔默(H. E. Palmer)、奥格登(G. K. Ogden)还有威斯特,都就词汇频率和外语学习词汇的有用性等问题展开了辩论。威斯特的最显著成果之一就是他 1935 年他编写成功的《新方法英语词典》(New Method English Dictionary)。

威斯特凭借自己丰富的课堂实践经验,以外语学习者的实际需求为准绳,严格限制词典的收词量,把它们局限在外语或第二语言学习者最有可能接触的范围之内,并尽当时所能,尽量收录当时的一些新词。[17]

其中最引人注目的就是威斯特在该词典的释义中只用了 1779 个单词,后来又减为 1490 个。威斯特控制词典释义词汇的目的在于减少学生的麻烦,其结果是写出的释义简洁明了。人们称赞威斯特的 NMED 具有划时代的意义,他创立了一种新的词典类型:英语教学词典。

在威斯特的影响下,后来陆续出版了第二代、第三代的英语单语学习词典。其中最为成功的例子是《朗曼当代英语词典》(Longman Dictionary of Contemporary English)(1978)。

"词典编写的最基本原则之一,是释义所使用的词语总是比被阐释的词语简单。"(《总论》)《朗曼》继承了个别词典试验过而中断多年的做法:"一切定义和用例所用词语被限制在两千个词左右,这些词语是在充分研究若干英语词汇频率表和教学用语表之后加以精选的。在这过程中,还特别参考了迈克尔·韦斯特的《英语一般词汇表》。"(《总论》)为了确保只使用两千个词的"中心"意义和较能为人所理解的派生词,编者采取了严格的措施,包括利用计算机进行检测,以及所有用例都出自编者手笔,而不拘于引证。[18]

《朗曼》的 2000 条释义词语广为人知,以致后来成为定义语言的代名词。那么这 2000 条释义词语是如何产生的呢?这主要靠的还是英语教师的语感,来源于经验。

可见,这里提取出来的定义语言是实际语言使用经验的产物,它因而必定具有常用、稳定、中性、基础、词义覆盖面广、现代性等特点。它们本身就是自然语言中的一部分。在

① 贺贤墚译,沈健凤、包含飞校《编辑和分类》,摘自《医学信息学手册》第六章,中国医药学术信息研究网站,http://www.cmia.org.cn/xsyj3/sanqi.html

② Michael P. West,有的著作译作"迈克尔·韦斯特"。

《朗曼》后来版本的变化中,虽有个别词语的增删,但其 2000 条释义词语的基本架构一直没有大的变化。

那么,这 2000 条词语是怎样的一些词语呢?1978 年版的《朗曼》后面附了释义用词表,经分析,发现通常说的 2000 条只是一个概数。准确地说是 2169 条,其中有前缀 13 条:dis - 、en - 、fore - 、im - 、in - 、ir - 、mid - 、mis - 、non - 、re - 、un - 、vice - 、well - ,有后缀 41 条:- able、- al、- an、- ance、- ar、- ate、- ation、- dom、- ed、- en、- ence、- er、- ess、- ful、- hood、- ible、- ic、- ical、- ing、- ion、- ish、- ist、- ity、- ive、- ization、- ize、- less、- like、- ly、- ment、- ness、- or、- ous、- ry、- ship、- th、- ure、- ward(s)、- work、- y、- ese。

在剩下的 2115 条中,有下面几种情况比较特殊:

A. 词义词性不同的同形词收入的有 7 组:bear(n),bear(v);lead(n),lead(v);March,march(v);May,may(v);Miss,miss(v);row(n),row(v);wind(n),wind(v)。

B. 收入意义相关词形稍有变化的两个词且并列同处一行的有 11 组:actor,actress;arch,archway;arrange,arrangement(s);Buddhist,- ism;child,children;Christian,Christianity;clothes,clothing;humour,humorous;sympathy,- etic;type,typical;violent,- ence。这些词并列作为一组是因为它们之间在词义与词形上有着密切的派生关系,其差异来源有的是性别的不同,有的是词性的不同,有的是单复数的不同,有的是本义与引申义的不同。如果严格地按词形不同则属不同词语的话,把它们拆开来分别排列,则会多出 11 个词。

C. 与上一类相似,也是意义与词形密切相关的词语,但不是以逗号隔开而是用括号标示放在原词后面的,共有 43 组。其中的原因除了上述各种外,还有一类就是固定搭配的结构:according(to)、affair(s)、alcohol(ic)、ash(es)、atom(ic)、attract(ive)、backward(s)、bacteria(- ium)、consonant(sound)、fashion(able)、forward(s)、gradual(ly)、her(s)、Hindu(ism)、indoor(s)、infect(ious)、jaw(s)、jealous(y)、Jew(ish)、lung(s)、moment(ary)、moral(s)、our(s)、outdoor(s)、ox(en)、plastic(s)、provision(s)、recent(y)、relative(s)、ruin(s)、scale(s)、scarce(ly)、sex(ual)、sock(s)、sport(s)、stair(s)、stocking(s)、their(s)、tropic(s)、vowel(sound)、worthy(of)、wrap(up)、your(s)。

D. 复合性词组的有 5 例:all right、god & God、no one、owing to、postage stamp。

为了更好地了解《朗曼》的释义词集的构成情况,把它与《朗文多功能分类词典》作了一个对比。《朗文》是目前英语学习词典中最受欢迎的分类词典之一。它的收词规模不大,但都是在学习英语时要求掌握的通用、常用、基本的词语,共收了以图示义和以文示义的词条 17061 条。原书末尾说"共收录词汇、词组、习惯用语近 30000 个",这里面其实是包括了许多在"立目"词语后面罗列的相关词条。《朗文》根据"人类的社会生活为中心,围绕着社会中的人"这一基本原则建构了语义分类系统,共分出 14 大类,129 中类和 2284 小类。由于《朗文》的这些性质,把《朗曼》释词与之对比,可以更好地看出《朗曼》释词的语义分布及词语选择的标准。① 为了方便在两个数据库之间进行联表查询,对《朗曼》2169 条释义略去前缀、后缀与同形词三类,其余的 2108 条,与《朗文》进行了对比,发现存在于《朗文》的有 2003 条,《朗曼》有而《朗文》无的有 105 条。②

《朗曼》2000 释词具有以下特点:

首先是通用性词语。《朗曼》释义词基本都在《朗文》所收的 17000 条词语的范围内,而后者正是供语言初学者使用的词典,它针对语言学习者而精选词语的做法是得到学术界公认的。定义语言的通用性,与学习词典的词语通用性,在这里正好吻合。其次是分布面广。《朗文》所收词语对整个英语词汇来说虽然不很充分,但在某种程度上来说却是完整的。因为要全面达到英语学习与使用的交际水平,它就必须照顾到语言表达与实际使用的各种需要。因此,所挑选出来的词语必须是全面的。而且《朗文》的分类正是体现出了一种语言的词汇整体使用功能的特点。以"人类的社会生活为中心,围绕着社会中的人",这正是人类社会,也就是语言社会的生成与分布的最大特点。《朗曼》的定义语言正好均匀地分布在语言社会的各个方面。

再次是常用性词语。《朗曼》定义语言的另一个特点就是常用性,必须通俗、易懂才能更好地完成对其它语言的说明与定义任务。在《朗文》中,同一小类的词语排列在前的词语大都就是常用词。如 A20"动物的幼仔"收了"young、offspring、progeny、issue、litter",《朗曼》释义词 young 位于其首。A54"狗与同类动物"收了"dog、puppy、hound、sheepdog、mongrel、cur、bitch",《朗曼》释义词 dog 位于其首。

最后是充分利用了英语词汇以派生构词法为主,构词能力强的优势,收录了数十个常用的构词缀,这样可以在有限的释词范围内较灵活地搭配重组,再造出新的适用词语来。但这种做法可能是太看重"2000"数字的封顶而不愿突破。把一些比词低一级的构词单位放到释义词语中,可视作是一种变通的做法。既做到了释义词语的"有限性",又不妨碍它

① 宋婧婧、苏新春《类义词典中的两种类型:"同义"与"同类"——〈同义词词林〉与〈朗文多功能分类词典〉比较》(2003),待刊。

② 这里说到的《朗文》中的词语是指"立目词语"。

的扩张性与组合能力。

长期以来,《朗曼》的有限释词的做法一直得到人们的赞同,2000 释义词也保持着相当的稳定性,与它以上的这些特点是分不开的。《朗曼》释义词语的选用很大程度借鉴了威斯特的词频统计结果,而威斯特的选词主要依靠的是长期的英语教学经验。结果证实,好的释义元语言一定要有很充实的语言使用经验作基础,也就是说要以"实用"为第一测量标准,而这正是语言最本质的特点——交际功能的实现。

四、张津、黄昌宁对汉语定义原语的提取

清华大学计算机系智能技术与系统国家重点实验室的张津、黄昌宁二先生 1996 年提出了国家自然科学基金重点项目研究报告《从单语词典中获取定义原语方法的研究及现代汉语定义原语的获取》,这是汉语学界第一份关于汉语释义元语言研究的报告。他们的做法是以有完整释义的汉语词典为封闭材料,通过数学模式来计算释义词与被释词之间的语义关系,从而得出最低数的释义词语,形成了含 3857 条词的释义元语言集。他们的工作是这样的:

1. 对象语料

使用的是《现代汉语词典》中所有复音词目的定义与《现代汉语通用字典》中所有单字词的定义。① 其所收录词语及义项容量如下:

	词型数	义项数	单义词数	多义词数	多义词平均义项
单字词	6517	15602	2997	3520	3.581
多字词	44386	52263	37822	6564	2.200
合计	50903	67865	40819	10084	2.682

2. 理论与方法

使用的数学理论和模型是"图论"。

根据这种形式化描述我们可以把词典转化成一种用有向图表示的方法,并为从词典中获取原语的问题建立一个数学模型,将这个问题转化为一个图论的问题。[6](P3)

图论是有着广泛应用领域的数学理论。它把研究对象都看作是一个有"顶点"和连接顶点的"边线"组成的"图"。图论就是研究这样的"顶点"与"边线"构成的"图"的科学,其实也就是研究"顶点"与"边线"的关系。"图"用 G 表示,V 表示"顶点",E 表示"边线",图论最简单的公式就是"G = [V, E]"。与 V 关联的边的条数叫做"顶点 V 的次数"。由于图中每条边都有两个不同的顶点,在计算次数时,每记一条边同时都要相应地记其端点两次(两个顶点)。"图论为任何一个包含了一种二元关系的系统提出了一个数学模型。"[19](P1)

图论在它的理论论述中有这样两点尤为值得我们注意:

1."一个'图'中各顶点之间相互位置的摆法一般来说对我们并不重要,我们关心于一个'图'的只是它具有哪些顶点;其次,图中所划出的各条边的长短、曲直我们也不予关心,我们关心的只是哪一个顶点与哪一个顶点之间有或者没有边相连。"

2."图论中不允许出现没有端点或只有一个端点的边。这是因为……一个图中的顶点可以代表各种事物,而边则代表事物之间的某种特定的关系,因而可以想像,不可能出现没有事物的'关系',也不可能只有一个事物而没有其对立面的'关系'。"[19](P21)

张、黄氏在图论原理下采取了下面的具体做法:

对于一部词典,我们可以用下面的方式将它唯一的对应于一个有向图。

1. 对于词集中的每个词 w2,在图中有且仅有一个节点 n 与它对应;② 且图中的每个节点 n,在词集中有且仅有一个词 w 与它对应。

2. 对于词集中的任意两个词 w1,w 2,(其中 w1 对应于 n1,w2 对应于 n2)。若 w1 的定义串中出现了 w2,则图中从 n1 到 n2 有一条有向边;否则,图中从 n1 到 n2 不存在有向边。[6](P6)

这样,充当释义词的机会愈多,其实就是充当"边"的端点的机会就愈多,也就有愈多的"边"集聚在它身上。转换成释义语言,也就是充当释义词的频率愈高,释义词的特征愈突出。这就是运用图论来提取释义元语言的最基本演算原理。应该说这是有很大合理性的。

3. 统计结果

经使用以上方法,共得出"定义原语"3856 条。研究者对这 3856 条释义词作了义类分布调查,参照依据是《同义词词林》(下面简称《词林》)。义类标注及分布调查成为研究者对统计结果所作的唯一加工及评价。研究者的说明是这样:

① 《现代汉语词典》,中国社会科学院语言研究所编,商务印书馆,1988 年第二版。《现代汉语通用字典》,傅兴岭主编,外语教学与研究出版社,1987 年。

② 这里的 n,相当于上文所引《图论基础教程》公式中的 V。——引用者注。

我们对定义原语集中原语义类分布进行了统计来观察由属于哪些义类的词或概念可以构成一个定义原语集。我们选用梅家驹等人主编的《同义词词林》中给出的义类代码。[6](P6)

这是汉语释义元语言首次"有形"地完整出现,无论从那个方面来说,它都是很值得让人们高兴的事情。不管它是清晰还是模糊,也不管它是漂亮还是丑陋,它的存在就说明一个新生命的存在,更何况是"新新人类"的第一个。当然,我们接下来的工作就是要评价它的生命力如何,评价导致它产生的生物基因如何。相信这一工作是很值得做的,因为这将使它的生命在更广阔的空间中得到延续。

在充分肯定这个研究结果的同时,还要看到它存在着若干明显的不足。对此本文做出以下四点评论。

其一,以《词林》作为义类标码的参照系是可以的,但释义词的有无以它作比较对象则显得大而无当。为什么这样说呢?《词林》是现代汉语的第一部义类词典,它把现代汉语词汇共分出了 12 大类,94 中类,1428 小类及 3927 小小类。虽然汉语分类词典后来还陆续出了一些,但分类之全面、妥贴,还是首推《词林》。这就是分类上以它为参照点可以的根据。又说它不可以,指的是它的收词范围。《词林》收词 7 万多,从篇幅上看与《现代汉语词典》、《当代汉语词典》、《新华词典》等现代的、中型的、语文性的词典收词大体相当或略大,但收词内容其实相当的杂芜。古今并收,雅俗同列,通方齐存,比比皆是。再从词汇单位来看,词与语、词与词素、词与短语,随处可见。因此,以统计出来的释义词见还是不见于《词林》,实在是没有什么参考价值。我们前面在说明《朗曼》2000 释义词特点时,把它与《朗文》作了对比,这是因为《朗文》的词语具有通用性、常用性、基础性等特点,而且全书的收词规模不大,精选程度高,而《词林》则离这样的要求相去甚远。

其二,对释义词资格的怀疑。上面我们说了《词林》的收词是多而滥,相当的杂芜。但即使是这样,3856 条释义词中仍有相当多的不见于《词林》:去掉重复的 58 组,① 不重复的词语是 3790 条,不见于《词林》的多达 1110 条。再与《现汉》对比,不见于其中的多达 797 条。这么多连如此规模的语文词典都不收录,它们怎么有资格成为释义元语言的一部分?下面我们来看看 3856 条释义词的语义分布。所有释词的义类分布见下表,并举出若干例词(接音序择选)。②

分类名	数量	例词
地理天文	169	阿拉伯海、巴尔干半岛、板块学说、半影、北斗星
电子、科技	72	半导体、半导体收音、变压器、步谈机、插头
法律	27	辩护、捕房、辞讼、从刑、大理寺、法定人数、法规
航空航天	9	第二宇宙速度、第一宇宙速度、飞机、飞行半径
经济	84	保护贸易、本币、剥削、不变价格、财务、财政寡头
军事	71	八路军、兵、兵役、兵种、驳壳枪、步兵
教育	31	班级、初等教育、初级小学、初级中学、传授、大学
建筑、交通	85	扳手、采莲船、叉车、铲土机、车、车钩、车轴
社会生活	147	哀子、救包、拔丝、白种、百事通、扳不倒儿
时间历史	111	奥陶纪、白垩纪、百日维新、拜上帝会、半夜、北朝
生命、医药	943	阿的平、阿米巴、阿米巴痢疾、阿尼林、阿片
体育	23	奥林匹克运动会、冰鞋、粉牌、滑冰、滑雪、象棋
物理、化学、数学	433	阿耳法粒子、阿耳法射线、爱克斯射线、安培
文艺影视	134	八股、八股文、伴奏、扮演、梆子、报告文学

① 重复的情况是:重出 4 次的有 1 组,重出 3 次的有 6 组,重出 2 次的有 51 组,共 73 条。

② 该研究的所有 3856 条释义词见于本文附录。限于篇幅,不刊。

新闻出版	20	报刊、报纸、出版、打字机、底片、电台、东昌纸
语文	1295	阿、安、安宁、安稳、安息、安逸、按摩、昂贵、凹
哲学、人、心理	106	安那其主义、八卦教、把斋、必然论、辩证法、财神
政治	96	安理会、八国联军、八字宪法、白皮书、百日王朝

下面再完整地穷一类之貌以窥其总概。属“军事”类的释词有 68 条词语:

八路军、兵、兵役、兵种、驳壳枪、步兵、陈胜吴广起义、冲锋、抽丁、初战、传爆线、弹壳、导弹、导弹快艇、第二次国内革命战争、第三次国内革命战争、第一次世界大战、短枪、法国大革命、反坦克炮、弓、海军、核武器、护卫艇、火焰喷射器、机关枪、甲午战争、歼击机、简师、局部战争、军队、军籍、军舰、军事、军衔、抗日战争、空降兵、来复线、榴霰弹、芦沟桥事变、炮弹、炮钎、炮塔、起义、枪弹、侵略、氢弹、燃烧弹、上校、生物武器、生物战、坦克兵、特种战争、铁甲车、退伍、无烟火药、武昌起义、武器、瞄炮、新四军、无后坐力炮、战列舰、战争、战争状态、中国工农红军、中将、中尉、作战。

“语文”类大体上都是通用词语、一般性词语,可它只有 1295 条,仅占总数的三分之一,非语文类词语却占 64.5%。而在“no-code”类中,非“语文”类的专有性名词则高达 89%。

其三,专有名词过多。对这个问题可从一个侧面,即词长来看出。这里的释义元语言的平均词长与《现汉》比较接近。下表是张、黄词语集的词长统计:

词长	1	2	3	4	5	6	7	≥8
词语数	557	2151	713	322	60	31	15	7
百分比	14.4%	55.8%	18.5%	8.4%	1.6%	0.8%	0.4%	0.2%

《现汉》(三版)的词长如下:

词长	1	2	3	4	5	6	7	≥8
词语数	10776	40207	4993	4852	218	106	48	61
百分比	17.6%	65.6%	8.2%	7.9%	0.4%	0.2%	0.1%	0.1%

前者的平均词长为 2.31,后者是 2.32。无论是音节长短的分布,还是平均词长,它的词长都比《现汉》略长。粗看去都还在可接受的范围,但落到细微处,特别是对较长的词语来作同类比较的话,则可以看出前者专有名词比例过大的特点还是相当的突出。请看双方长词长的比较。这是张、黄统计中的 ≥7 的 22 条词语(按音序排列):

奥林匹克运动会、博斯普鲁斯海峡、超外差式收音机、第二次国内革命战争、第三次国内革命战争、第一次世界大战、各尽所能按劳分配、各尽所能按需分配、共产主义青年团、共产主义者同盟、古典政治经济学、广西壮族自治区、和平共处五项原则、流行性脑脊髓膜炎、流行性乙型脑炎、马克思列宁主义、马克思主义哲学、美尼尔氏综合症、朴素的唯物主义、人民代表大会制、社会主义所有制、新疆维吾尔自治区。

而《现汉》中词长为 7 的词语有 48 条,大于 7 的有 61 条。下面按音序排列的前 20 条:

按下葫芦浮起瓢、盎格鲁撒克逊人、奥林匹克运动会、百尺竿头更进一步、百花齐放百家争鸣、百足之虫死而不僵、搬起石头打自己的脚、饱汉不知饿汉饥、冰冻三尺非一日之寒、兵来将挡水来土掩、不到黄河心不死、不管三七二十一、不见棺材不落泪、不经一事不长一智、不入虎穴焉得虎子、差之毫厘谬之千里、长江后浪推前浪、车到山前必有路、成事不足败事有余、成也萧何败也萧何。

两相对比,前者专有名词的比例之大是显而易见,这样必然会影响到“释义元语言”的功能。研究者对此也有所察觉,“这样就导致了一些专有名词收录到定义原语集中”。但问题在于,这里的“一些”竟达到了总数的三分之一。对“释义元语言集”的研究来说,这是一个难以容忍的巨大数字。更重要的是,研究者对此显得过于宽容,没有作任何的甄别筛选,仅仅是因为技术上的原因,因为“词典定义的每个环路上都至少要有一个词被收录到定义原语集中。由于这两个词之间的定义出现循环,这两个词中至少有一个要被收入定义原语集中”[6](P24),就对大量专有名词进入元语言的行列采取了默认的态度,这其实是唯技术、唯方法的结果。完全依赖于所使用的数学方法,把异常复杂的语言问题简而化之,这显然是对元语言性质认定的复杂性及提取工作的繁难性缺乏足够的认识。

其四,对 3856 条释义词的词性进行调查,发现名词高达

2865 条,超过七成。名词比率过高是专有名词过多的当然反映。从一般的分布规则来看,在最重要的、最基本的词语中,名、动、形三类词语之间会保有一定的比例,名词会多些,但动词与形容词也属于最重要的基本词类,所占数量不应太少。愈是在重要的、基本的词汇中,三者的比例差距愈不会拉得太开。一般会在 5∶3∶2 之间。这点我们可以从许多基础性的材料看出一个大概。

如对外汉语教学用的《汉语水平词汇等级大纲》收录了 8822 条词语。这属于基本词、常用词。标注词性的有 8145 条,未标注词性的 677 条。标注词性的有名词 3581,动词 2774,形容词 1077,其它 11 类 713,比率为 4.3∶3.4∶1.3∶0.8。[20]

《朗文》所收的 17000 余条词语,属英语常用词,99.7 %的词语标注了词性。除掉兼类,单一词性的 15360 条,其中名词 9808,动词 3110,形容词 2194,其它 248,比率是 6.3∶2∶1.4∶0.1。

从理论上看,释义元语言既不属于专书专人的个人言语作品,也不像基本词汇那样处于词汇的底层,但它的常用词、通用词、高频词等特性却是显然的,专有名词不可能那么多也是显而易见的。

张、黄研究的学术价值是显然的。一是它所利用的材料具有从事"释义元语言"研究的现实意义,是以真实的词典材料为蓝本来从中提取,所选取的语料也具有相当高的权威性和代表性;二是它属于完全的计算语言学范畴的工作,依靠对语义关系的算法来确定释词的地位;三是最终形成了具体有形有数的汉语释义元语言的词汇集。这与语义哲学中的元语言研究有着很大的不同。语义哲学讲究的是形式语言的单位与表达格式,而非真实语言的提取;是语言的逻辑生成而非语言的释义表达;是若干语言指称符号与表达规则而非释义词语的完整词集。特别是它属对汉语释义元语言的首次提取,筚路蓝缕,功莫大焉,它将大大启发人们在这条道路上继续走下去,人们都将会记住这一开创性的研究。

但它也存在若干不足。最突出的就是提取出来的释义词可信度不高,里面存在大量不可能承担起释义功能的词语。其原因很大程度就在于提取方法的选择。理论上说,数学图论方法是正确的,但实际效果并不好。原因何在?关键可能就在于对"顶点"的认识与取舍上。由于环路上的"结点"对象,其中必有一个成为收录对象,而专有名词在释义上又是最不容易继续往下贯通的,因此也就大量地被收录了进来。而理想的释义词恰恰是具有连续释义、"左右逢源"的特点。看来在研究释义元语言时,不能单纯地站在"理论"与"方法"的角度来观察释义词,还应该从"释义词"的角度来反观"理论"与"方法"。这应该是选取、运用、评价"理论"与"方法"最重要的视角。

正是这点启发我们做出这样的思考:提取释义元语言还有没有别的理论和方法可以利用?用怎样的理论和方法更符合释义元语言的本质特点与功能?这是正确的释义型元语言研究必须首先解决的问题。

五、国内其它的汉语元语言研究介绍

当前国内语言学界对元语言的研究兴趣正在加强,其中最值得注意的是南京师大李葆嘉的研究。2001 年 12 月他与安华林合作撰写论文《论现代汉语元语言系统研究》,提交给在厦门大学举行的"全国汉语词汇规范研讨会",首次展示了他关于这一研究的看法。① 他认为:"语言学元语言包含三层含义:用于语言交际的最低限量的日常词汇,用于辞书编纂和语言教学的释义元语言,用于语义特征分析的析义元语言。"不同类型的研究所得出的元语言单位是不同的。"元语言研究的对象应当包括'语元单位'(简称'语元')和'语元关系'(语元间的语义关联)。词汇元语言中的语元相当于日常语言的基本词项,释义元语言中的语元相当于词典释义中的必用词项,现统称之为'词元'。析义元语言中的语元单位,通常用义素、语义特征、义征、语义元语、语义标示语、义元等术语表述,现与机用元语言系统中的语元单位统称之为'义元'。认知元语言中的语元单位即认知元,则称之为'知元'。"从而也可以得出三种不同类型的元语言集:释义元语言系统、最低限量词汇集、语义标记集。

后来他对这一问题作了进一步的思考,② 并愈来愈倾向于从自然语言处理的角度来思考问题。他最近完成的一篇论文是《人工语言学脑:自然语言处理装置的研制思路》。③ 认为"建构元语言系统、提取和标注义征的微观分析过程,就是'语言基因图谱分析工程'"。其目的是要建造由计算机来运行的"人工脑",包括:"作为研制基础的'语言基因图谱分析工程',作为研制关键的'认知语义网路建构工程'和作为研制目标的'受限语言能力模拟工程'。"现在还没有具体看到李氏对元语言分析、提取的具体结果,但他关于

① 论文略加修改,收入苏新春、苏宝荣主编《词汇学理论与实践(之二)》,商务印书馆,2003 年 7 月。

② 见李葆嘉《汉语元语言系统研究的理论建构及其应用价值》(《南京师范大学学报》2002 年第 4 期)、《语义语法学理论和元语言系统研究》(《深圳大学学报》2003 年第 2 期)。

③ 李葆嘉《人工语言学脑:自然语言处理装置的研制思路》,提交给由香港城市大学主持召开的"第四届汉语词汇语义学研讨会"(网上研讨会),2003 年 7 月。

元语言研究理论层面的诸多考虑,特别是关于计算机自然语言处理方面的意见,是富有建设性的。

参考文献:

[1]王路.走进分析[M].上海:三联书店,1999.
[2]黄玉顺.语言的牢笼——西方哲学根本传统的一种阐明[DB/OL]."思问"哲学网站,http://www.siwen.org/renshilun2003-07-09.
[3]卡尔·波普尔.无尽的探索——卡尔·波普尔自传[DB/OL],邱仁宗译,远大教育网站,http://city.yondor.com/library/philosaphy/b/bopuer/wjdt/033.htm
[4]R.R.K.哈特曼,R.C.斯托克.语言与语言学词典[M].黄长著等译,上海:上海辞书出版社,1981.
[5]戚雨村,等.语言学百科词典[M].上海:上海辞书出版社,1993.
[6]张津,黄昌宁.从单语词典中获取定义原语方法的研究及现代汉语定义原语的获取[R].1996.
[7]张志毅,张庆云.词汇语义学[M].北京:商务印书馆,2001.
[8]黄建华.词典论[M].上海:上海辞书出版社,2001.
[9]《当代中国心理学》赏析:汉语句子加工过程和儿童阅读发展研究[DB/OL].人民教育出版社网站,http://www.pep.com.cn2003-07-09.
[10]钟启泉.我们的中小学生需要怎样的语文素养[DB/OL].人民教育出版社网站,http://www.pep.com.cn/2002-12/ca11005.htm
[11]舒华.汉语儿童语音意识的发展[DB/OL].http://www.bku.cn/cscsl-y/newwoks/Groth 2003-07-09.
[12]李子荣.漫谈盛可以说语言艺术(上)[DB/OL].新浪读书,http://book.sina.com.cn/2003-06-05/3/8439.shtml
[13]XML 与电子病历[DB/OL].三九健康网,http://fm365.39.net.cn/professional/medicine/200208/13358920020829.htm
[14]揭开 XML 的神秘面纱[DB/OL].新电脑杂志社,http://www.chip-china.com/1006.phy?sid=1143.
[15]藏策.批评 元语言 文化研究[DB/OL].中国摄影家协会网.
[16]于晖.语篇体裁结构潜势及其应[DB/OL].中国翻译网,http://www.chinatranslate.net/asp/gb/tradewindows/wind-read.asp?key=119.
[17]张利伟.英语学习词典的起源、发展和影响[J].外语教学与研究,1996,(3).
[18]陈炳超.评《朗曼当代英语词典》[J].辞书研究,1982,(3).
[19]赵宏量,彭太华.图论基础教程[M].重庆:西南师范大学出版社,1988.
[20]国家对外汉语教学领导小组办公室.汉语水平词汇与汉字等级大纲[Z].北京:北京语言学院出版社,1992.
[21]李葆嘉.汉语元语言系统研究的理论建构及其应用价值[J].南京师范大学学报,2002,(4).

(责任编辑:言 之)